International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

# CUSTOM NAMED ENTITY RECOGNITION FROM CORPUS DATA USING CONDITIONAL RANDOM FIELD

[1]M Eliazer, [2]Parvathy S, [3]Akshara Santharam, [4]Biswas Sreya Monobikash

***ABSTRACT--****Finance is a growing field which is concerned with the allocation of assets and liabilities over space and time. In any wealth management organization, several advise set documents are used where financial statements and client data are recorded. Therefore, the identification and recognition of custom named entities using Natural Language Processing could help the clients to understand and find relevant information from the extracted data. A new model using the method Conditional Random Field (CRF) is developed as it provides accurate results when compared to already existing model. The proposed model is fast, highly accurate, easy to install and use. This paper proposes a method to extract custom named entities from corpus data of financial domain using Conditional Random Field (CRF) and evaluates the effectiveness of the proposed method.*

***Key words--*** *Domain Specific Entity, Named Entity Recognition, Conditional Random Field (CRF), Word Embeddings, Extraction, Wealth Management, Advise set documents, Custom Entities.*

## I. INTRODUCTION

Named entity recognition (NER) is also called as entity extraction or entity chunking and is a major task in Natural Language Processing [**1**]. It is been widely used in all areas and research including medicine, finance, banking etc. Different entities like Person, Organization, Location, Places, etc. can be recognized and extracted using Named Entity Recognition approaches. Example, Ram is from Chennai, in this Ram and Chennai is named substances where a Named Entity Recognition (NER) system must identify RAM as a 'name of the person' and CHENNAI as the 'name of a place'. The basic steps involved in identifying an entity from a raw text is Sentence Segmentation, Tokenization, Part-of-Speech (POS) Tagging, Entity Recognition and Extraction [**2**].

Finance Companies deal with a huge quantity of advice set documents which is acquired from the clients, like paper documents, digital reports etc. Advice set documents contain a lot of information about the clients, services offered and clients requirements. Therefore, we have unstructured data to be processed by the natural language processing approach like named entity recognition.

*[1]Assistant Professor, Dept. of Computer Science Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India , eliazer.m@ktr.srmuniv.ac.in*

*[2]Student, Dept. of Computer Science Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India, psp2397@gmail.com*

*[3]Student, Dept. of Computer Science Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India, vaidehisantharam@gmail.com*

*[4]Student, Dept. of Computer Science Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India, sreyabiswasxyz12@gmail.com*

**Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020** 

The rest of the article is organized as follows: In section 2 we described related work about Named Entity Recognition, Information Extraction and Natural Language Processing. In section 3 problem statement is been presented along with our solution in section 4. In section 5, the method that can be used to extract entities required, which includes the training of data set, pipeline and training the recognizer. In section 6, our proposed model is being mentioned Additionally, in section 7 we presented the experimental setup where the number of entities extracted and other results that we obtain is being stated (section 8) along with the tools employed and the processing pipeline. The paper ends with conclusion and future work in section 8.

## II. RELATED WORK

Shamima Parvez et al. [**3**], focuses on the improvement of Bengali Named Entity Recognition system. It lacks annotated data as well as has a very less accurate POS tagger. The system preferred to try to overcome this in three steps. A new POS tagger is created, and data is trained and tested using HMM technique. Still the model did not provide a good accuracy and limitation in the availability of corpus was a great challenge. The future work suggests in developing a model that can accurately do the entity extraction process.

Waleed Zaghloul et al. [**4**], proposed an entity extractor with high precision, that can extract entities like person and organization. Almost 52,000 data sets were used to identify both these entities using a Rule-based entity extractor. Different attributes like precision, accuracy and quality of the extractor is checked. The paper suggests making a good extractor with high precision and accuracy, that can work well with all knowledge domain.

Payal Biswas et al. [**5**], proposes a Named Entity Recognition system called AGNER that can be mainly used in the agricultural domain. The main process involves domain specific tagging and linguistic processing. The paper concludes by creating an entity extractor that works well with requirements of agricultural domain.

Dan Klein et al. [**6**], depicts the importance of character and character n-gram in Named Entity Recognition models. Conditional Markov model and character level HMM model is being used. The

paper concludes but stating that switching from word model to character model can produce a greater accuracy of about 30%, also it is prone to less error.

Vivek Kulkarni et al. [**7**], discuss method to analyse the semantic difference that exist in word usage. Methods to identify semantic of word usage and variations in linguistics is analysed. Methods like Domain specific word embedding and word2vec is being employed. The paper concludes by identifying all the domain specific entities and works well with large corpus of data.

Dr. K.S.Wagh et al. [**8**], uses Conditional Random Field and machine learning techniques to extract the entities. In this paper, terms like DNA, RNA, cell line and type, protein etc are being extracted. The paper concludes by precisely extracting all the required biomedical entities and terms using the specified machine learning techniques.

Ken Yano et al. [**9**], uses recurrent neural network along with bi-directional LSTM coupled with Conditional Random Field. In this paper, they try to extract disease name as entities using a Japanese Medical Text. The paper concludes by effectively finding a method to extract the entities from the Japanese text using simple processing and automatic learning about the Disease Named Entity (DNE) Extraction.

Meizhi Ju et al. [**10**], proposes a novel method for entity recognition using dynamic neural model. The two data sets used here are GENIA and ACE2005.The paper compares the performance of dynamic neural method and

traditional method and came to conclusion that proposed method works well that the existing one. Also, the proposed method worked more accurately with both the data sets even when it contained nested entities.

Shuwei Wang et al. [**11**], proposed a method to identify entities from Financial domain. The domain dictionary is used first, so that the system can recognize the financial terms. Then the abbreviations of the financial terms are identified. Mutual information boundary is being used for this purpose. The paper concludes by stating the approach that successfully improved the accuracy and helped in easy extraction of financial named entities.

Hutchatai Chanlekha et al. [**12**], proposes a method to do Thai Named Entity Recognition. Simple heuristic information along with maximum entropy model is being employed for this purpose. The data set used is a huge corpus with almost 110,000 words for training and 25,000 words for testing in the political domain. The research concludes by stating the proposed method as acceptable and suggests using more powerful methods in the future for the entity extraction.

## III. PROBLEM STATEMENT

The major issue with the existing model is disambiguation i.e., the confusions about whether the word in a text corpus denotes the name of a person, a location, an organization etc. Another issue is that the existing model is not suitable for large sets of data consisting of thousands of advices set documents. Also, the precision and quality of extracted data is poor in such models. A lot of time is taken for extracting and identifying the entities. Also, the existing models cannot extract Custom Entities from the corpus data.

## IV. SOLUTION

We must identify and extract domain-specific entities from the advice set documents that represent all supporting documents provided by different agents to Wealth management end

clients. The main objective is to identify custom entities, as per the requirement of the domain. Example, “Virat Kohli is the captain of Indian National Cricket Team”, in this ‘Virat Kohli’ is to be identified as a ‘Cricketer’ rather than as a ‘Person’. To achieve this and overcome the existing limitations, various Natural Language Processing (NLP) approaches are used.

## V. METHOD FOR EXTRACTING CUSTOM ENTITIES

In this study, we focus on that extracting custom entities is one of the tasks of Named Entity Recognition. Therefore, we propose a method for extracting custom entities from advice set documents by Named Entity Recognition in this paper. The proposed paper uses Conditional Random Field (CRF) to Named Entity Recognition. In this paper, the extraction target is only entities in Financial Domain like the name of financial organizations, services offered etc.

### *5.1 Training the Dataset*

Algorithms learn from data. They find relationships, develop understanding, make decisions, and evaluate their confidence from the training data they're given. Better the training data is, the better the model performs. To accomplish our goal of achieving a highly accurate model, we plan to use a vast annotated API, which would be used as training data, that would not just help to identify the correct label for the testing data set entity on, but also establish a relationship between entities.

### *5.2 Building the Pipeline*

To create our own entity recognizer, we need to create a pipeline, for which we are loading a blank model in the English language with the help of spacy tools. The purpose of the pipeline is to load the interface where we train our recognizer to scan through the test data input by a user and label their entities accordingly.

### *5.3 Training the Recognizer*

This step happens to be a very crucial section of the entire process. In this step, we are training the recognizer in such a manner that the recognizer can identify and label a test data that has been given by the input based on the knowledge it has learned from the training data set. This step involves random shuffling the training data so that the recognizer can establish a relationship between entities and in turn increase its accuracy of labelling. Once done so, all other modules of the pipeline need to be disabled so that only our Named Entity Recognition model is being executed.

### *5.4 Testing the Trained Model*

Once the data set has been completely annotated and the recognizer has been trained upon it correctly, we need to make sure that the recognizer is indeed working well, and not generating any error, thus testing plays a major role towards the end of the project. If the recognizer is generating any error, we need to rework on the training set and recognizer to see to it that the error is eliminated and is recognizing the entities accurately.

### *5.5 Saving the Trained Model*

Once the recognizer has been trained properly on the training data set, and is performing as desired, we need to save the model so that it can be used as and when required and we then test the saved model.

## VI. PROPOSED METHOD

In this study, we create a new model for Named Entity Recognition (NER) algorithms, where the model requires a user to upload advice set documents and add the labels. The model then identifies attributes and extracts its corresponding values. The model relies on an NLP approaches to identify domain specific relationships. The entities that we are extracting are name of the Financial organizations, name of government organizations and different services offered. We use tools like Brat and SPACY for the annotation and entity extraction purposes. Open source possibilities including the Lucene Segmenting Tokenizer and the Open NLP sentence and paragraph boundary detectors are used

to identify and mark sentence, phrase, and paragraph boundaries. This project will expand the capabilities of existing model and build a model that can work with larger data sets and hence produce a more accurate result.

# VII. EXPERIMENTAL SETUP

In the following paragraphs, we have described the setup procedure for framework evaluation, the datasets and the frameworks used.

### *7.1 Setup Procedure*

We have made a setup procedure as shown in figure below (**Fig. 1**), to explain the steps taken to extract custom entities from corpus data. In step one, we use advise set documents as the input dataset that contains financial statements of the clients. The advice documents in MS Word Format was converted to JSON format and is used in the Spacy pipeline. In step two, the dataset extracted will be processed by the Named Entity Recognition modules using different tools like SPACY, BRAT etc. In the last step, custom entities are recognized and extracted.

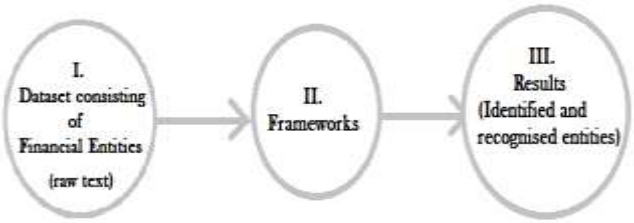


**Fig. 1:** Procedure Workflow

### *7.2. Dataset*

We used advice set documents as a data set which is provided by the finance company. The original file is in Microsoft Word Format with the following properties:

-Word count: 671

-Character count: 12144

-Line count: 492

-Paragraph count: 100

The original file is converted into a JSON format and entities are annotated using BRAT annotation tool. The focus of the experiment is to detect custom entities., e.g.., the name of the financial organization (FINANCIAL ORG), Government organization (GOVT. ORG), services (SERV) etc. The rules followed are:

- Financial organizations (FINANCIAL ORG): Identify the name of the financial organization with a minimum of 5 words.
- Government organization (GOVT. ORG): Identify the name of the govt. An organization with a minimum of 5 words.
- Services (SERV): Identify the services offered by financial organizations.

Entities extracted from our annotation procedure produce the following values: Financial organizations (FINANCIAL ORG)-489; Government organization (GOVT. ORG)-51; Services (SERV) -121.

### *7.3. Tools Used*

#### *7.3.1 Stanford NER*

Another popular name by which it goes is the Conditional Random Field (CRF) Classifier. This software makes use of a sequence model that implements the technique of linear chain Conditional Random Field. This technique is used to build a highly accurate entity recognition model.

### *7.3.2 Spacy*

Spacy is a not-so-old open source library that performs industrial-strength advanced natural language processing activities using python language on a large corpus at a more rapid speed. The named entity recognizer model provided by Spacy has been trained on the Onto Notes 5 corpus and can support many entity types. Apart from this, it also enables user to create its own entity type.

### *7.3.3 NLTK*

NLTK otherwise known as the Natural Language Toolkit is a Python package that provides a rather large set of natural languages corpora and APIs of varieties of Natural Language Processing algorithms. To perform the Named Entity Recognition task using the NLTK technique, three stages of tasks must be fulfilled: Work Tokenization, Parts of Speech (POS) tagging and Named Entity Recognition.

### *7.3.4 BRAT*

Brat is a web-based tool that makes it easy to annotate a large dataset and add notes to it. It is particularly designed for structured annotation where notes aren't free from text, instead they exist in a format that can be automatically processed or by computer.

### *7.3.5 Prodigy*

Prodigy is another brilliant annotation tool that supports active learning, so that once an entity has been allotted a label, it does not again question you about that entity, thus making the annotating procedure of the named entity recognition procedure much faster.

### *7.4 Pipeline*

When the processing model is called upon the test document, the latter gets processed in several steps, these steps combined are referred to as the processing pipeline. Each component in the pipeline processes the document and passes it on to the next component of the pipeline. The first step involves adding the labels (FINANCIAL ORG., GOVT. ORG., SERV to the pipeline) and loading the processing model. It is then followed by initializing the custom recognizer model that we have developed so that if there is any other model that already exists in the system, it gets nullified.

**Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020** 

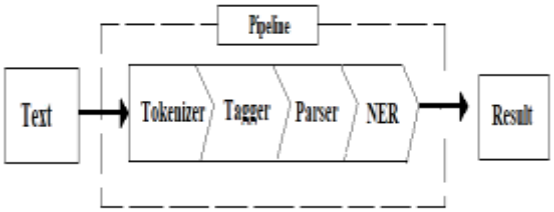


**Fig. 2:** Processing Pipeline

Finally, we begin training the data so that when the test data is inputted by the user or customer, the desired entities get labelled correctly. When the user inputs the data into the model, the data first gets tokenized to separate words which are then tagged and parsed inside the pipeline before it is made to go through the customized named entity recognition model (NER) that we have developed to correctly recognize and label the entities.

## VIII. RESULT AND ANALYSIS

The outcome we intend to obtain is to create an extractor that can recognize and extract custom entities as per the requirement of the domain. A custom entity recognizer that works well with large corpus data as well as minimizes the time consumed to extract entities. By using software library like Spacy, it is possible to extract custom entities from a large corpus

| Financial Advice set corpus | |
|---|---|
| Sentences | 492 |
| Tokens | 12144 |
| Named-Entities | |
| FINANCIAL ORG | 489 |
| GOVT. ORG | 51 |
| SERV | 121 |

**Table 1:** Data Summary

Following the setup procedure, we evaluated our dataset with the proposed method and obtained the result as shown in table 2:

| Entities Extracted | Precision |
|---|---|
| Financial Organization | 97 % |
| Government Organization | 95 % |

**Table 2:** Evaluation results

On comparing with the other models, our model (using the financial data set) has produced an overall accuracy of 96%, for the detected entities like Financial Organization and Government Organization. There was an entity that was not detected using our corpus, the Service entity and this is because of the unavailability of enough data for a fine train.

## IX. CONCLUSION AND FUTURE WORK

In this paper, we have proposed the method to extract custom named entities from the advice set documents using Conditional Random Field. The model developed has overcome all the limitations of existing models and has given an accurate and precise result. We were able to extract Custom entities as per the requirement of the domain and obtained promising results using the tools like BRAT and Spacy.

Future work will consist of framework improvement and developing a model that can work well for all the domain. A model to effectively do entity relation can also be developed. Finally, we will construct a system for extracting useful information about the entity.

## X. COMMENTS

**10.1** Reviewer **1: Dr. R Annie Uthra**

**10.1.1** The purpose of using the Conditional Random Field technique and the precision that is achieved in comparison to the other methods were asked.

**10.1.2** The annotation technique and its use for training the dataset were asked.

**10.2** Reviewer **2: Ms. K R Jansi**

**10.2.1** The format in which the user will be uploading the data into the model was asked.

## REFERENCES

1. https://en.wikipedia.org/wiki/Named-entity_recognition
2. https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da
3. Shamima Parvez-Named Entity Recognition from Bengali Newspaper Data- International Journal on Natural Language Computing (IJNLC) Vol. 6, No.3, June 2017
4. Waleed Zaghloul and Silvana Trimi: Developing an innovative entity extraction method for unstructured data - Zaghloul and Trimi International Journal of Quality Innovation (2017)-3:3-Published online -22 May 2017
5. Payal Biswas, Aditi Sharan, and Ashish Kumar AGNER: Entity Tagger in Agricultural Domain-2015 2nd International Conference on Computing for Sustainable Global Development (INDIA Com) 2015-IEEE paper
6. Dan Klien, Joseph Smarr, Huy Nguyen, Christopher D Manning- Named Entity Recognition with character-level models
7. Vivek Kulkarni, Yashar Mehdad and Troy Chevalier - Domain Adaptation for NER in online media with word embeddings- arXiv:1612.00148v1 [cs.CL] 1 Dec 2016

**Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020** 

8. Dr. K.S.Wagh, Aishwarya Kulkarni, Shraddha Kashid, Neha Kirange, Pratiksha Pawar-CRF based Bio-Medical Named Entity Recognition-International Journal of Emerging Technology and Computer Science-Volume: 3 Issue: 2 April – 2018;14-18
9. Ken Yano Neural Disease Named Entity Extraction with Character-based BiLSTM+ CRF in Japanese Medical Tex-arXiv:1806.03648v1 [cs.CL] 10 Jun 2018
10. Meizhi Ju, Makoto Miwa and Sophia Ananiadou -A Neural Layered Model for Nested Named Entity Recognition-Proceedings of NAACL-HLT 2018, pages 1446–1459
11. Shuwei Wang, Ruifeng Xu, Bin Liu, Lin Gui, and Yu Zhou- Financial NER based on conditional random fields and information entropy-Proceedings of 2014 International Conference on Machine Learning and Cybernetics, Lanzhou,13-16, July 2014
12. Hutchatai Chanlekha, Asanee Kawtrakul- Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information